基于 SCDV 及各向异性调整 BERT 的文本语义消歧方法 *

李保珍，顾秀莲
(南京审计大学信息工程学院，南京 211815)

摘　要： 文本表示需要解决文本词语的歧义性问题，并能够准确界定词语在特定上下文语境中的语义特征。针对词语的多义性及语境特征问题，提出了一种文本语义消歧的 SCDVAB 模型。主要创新点有：基于分区平均技术，将场景语料库转换为文档嵌入，并引入各向异性，改进了软聚类的稀疏复合文档向量(SCDV)算法，以提高 BERT 的语境化表示能力；将调整各向异性后的 BERT 词语嵌入，作为静态词语向量的文档嵌入，以提升文本语义消歧的能力。通过大量实验进一步证明，SCDVAB 模型的效果明显优于传统的文本消歧算法， SCDVAB 模型可有效提高文本语义消歧的综合性能。
关键词： 语义消歧；各向异性；BERT；稀疏复合文档向量；文本表示
中图分类号： TP391　　doi: 10.19734/j.issn.1001-3695.2022.03.0094

Text semantic disambiguation based on SCDV and anisotropy adjusted BERT

Li Baozhen, Gu Xiulian
(*College of Information Engineering*, *Nanjing Audit University*, *Nanjing* 211815, *China*)

Abstract: Solving the problem of ambiguity of text words is important for text representation, and it can accurately define the semantic characteristics of words in a specific context. Aiming at the polysemy and contextual characteristics of words, this paper proposed a semantic disambiguation model of SCDVAB. The main innovations are: Based on the partition average technology, it can convert scene corpus into document embedding, and introduce anisotropy to improve the sparse composite document vector (SCDV) algorithm of soft clustering to improve the contextual representation ability of BERT; and then it can improve the ability of text semantic disambiguation by embedding the BERT words after adjusting the anisotropy as a static word vector. Through many experiments, SCDVAB model is significantly better than the traditional text disambiguation algorithm. SCDVAB model can effectively improve the comprehensive performance of text semantic disambiguation.
Key words: semantic disambiguation; anisotropy; bidirectional encoder representations from transformers (BERT); SCDV; text representation

0　引言

文本语义高度依赖于组成文本的词语，同一词语在不同的上下文语境中，可能具有不同的含义，进而存在歧义性干扰。如何通过消歧来提高文本表示的准确性，一直是理论和实践所关注的重点。 对文本表示的一系列研究表明，用于句子表示的词向量加权平均通常优于更复杂的神经模型。SCDV(Sparse Composite Document Vectors，稀疏复合文档向量)将能够界定词语场景性语义的词语嵌入模型与能够处理不同词义的潜在主题模型结合起来，可增强词语的表达能力。使用嵌入的软聚类技术，可有效学习主题特征空间，通过文档向量的稀疏化操作，可减少处理向量任务的时间和空间复杂性，并能够有效处理文本表示的分布式段落向量。

静态词嵌入的一个显著问题是多义词的所有含义共用一个固定的静态向量，但静态词向量难以有效解决一词多义问题。用基于上下文语境的词语嵌入代替静态词嵌入可以提高词语消歧的效果，如以 BERT 为例的深层神经语言模型可将静态嵌入替换为上下文语境的词嵌入。通过预训练的 BERT 模型能够将多义词分别放置在具有不同含义的语义空间中，进而可输出不同的词向量，可解决静态嵌入无法有效解决的一词多义问题，实现基于语境化嵌入的可解释词义消歧。此外，BERT 模型中上下文语境性词语表示具有各向异性的特征，即它们在不同方向上不是均匀分布的，在向量空间中占据一个狭窄的圆锥体 [1]。各向异性是指文本词语的全部或部分含义随着语义空间维度方向的改变而有所变化，在不同的语义空间维度方向上呈现出差异的性质。例如词语“苹果”，在上下文语境为水果的语义空间维度中，在水果相关特征维度的方向具有更为显著的投影；在上下文语境为电子产品的语义空间维度中，在电子产品相关特征维度方向具有更为显著的投影。一个词语的语境化表示中只有不到 5%的差异可以用该词语的静态嵌入来解释[2]。这也为调整各向异性，减少各向异性对文本词语语境化表示的影响提供了必要性理由。

针对上述问题，本文提出一种简单有效的无监督表示方法 SCDVAB(SCDV+Anisotrop+BERT)模型。主要创新点为：a)通过软聚类的稀疏复合文档向量(SCDV)分区平均技术，将场景语料库转换为文档嵌入；b)在 SCDV 流程中，基于自相似性、句内相似性及最大可解释方差调整各向异性，以提高 BERT 的语境化表示能力；c)将调整各向异性后的 BERT 词语嵌入作为静态词语向量的文档嵌入，以提升文本语义消歧的能力。相关实验结果显示出 SCDVAB 模型在精确性上优于现有技术，能够提高概念匹配及语义文本相似度等相关任务的性能。

1　相关工作

对于短文本和文档表示任务，需要将词语嵌入扩展到整

收稿日期：2022-03-05；修回日期：2022-04-28　基金项目：国家自然科学基金资助项目(72074117，71673122)；江苏现代财税治理协同创新中心资助项目(20WTB007)；江苏省研究生科研创新项目(KYCX21_1948)
作者简介：李保珍(1975-)，男，山西晋中人，教授，硕导，博士，主要研究方向为网络大数据分析、文本挖掘(bzli@nau.edu.cn)；顾秀莲(1997-)，女，江苏盐城人，硕士，主要研究方向为自然语言处理.

个段落和文档。Le 和 Mikolov 在 2014 年提出了两种文本分布式表示模型，即分布式内存模型段落向量(PV-DM)和分布式 BoWs 段落向量(PV-DBoW)，将每个句子视为共享的全局潜在向量[3]。这两种模型训练词语和段落向量来预测上下文，但在段落之间共享词嵌入。然而，词语在不同的语境中可能有不同的语义。在包含相同词语的两个不同意义上的文本的向量需要考虑这种区别，以便准确地表示文本的语义。此外，尽管段落向量可以包含多个主题和多个词义，但它与词语向量嵌入在同一空间中。段落向量还假设所有词语在权重和质量上的贡献相等，这忽略了词语在不同文本中的重要性和独特性。

Ling 将词语嵌入映射到潜在主题空间，以捕捉词语出现的不同意义[4]。但是,在与文字相同的空间中表示复杂文档，降低了表达能力。2015 年，Mukerjee 等人提出了词语向量的 idf 加权平均，以形成文档向量[5]。但是，其假定文本中的所有词语都属于同一语义主题。Gupta 在 2016 年提出了一种使用词语嵌入和 tf-idf 值形成复合文档向量的方法，称为词语包向量(BoWV)[6]。BoWV 背后的核心思想是语义不同的词属于不同的主题，但是该模型的词向量平均设置具有一定的局限性。

Mekala 等人在 2017 年通过对预先计算的词向量进行软稀疏聚类，使用 tf-idf 加权形成稀疏复合文档向量(SCDV)[7]。SCDV 作为一种文档的特征向量形成技术克服了广泛用于文本表示的分布式段落向量表示的一些缺点。然而这种方法在一定程度上忽略了文本词语的歧义性问题以及上下文语境的语义特征问题。2020 年，Gupta 等人在字向量上获得的多感嵌入将 SCDV 扩展到了 SCDV-MS，强调了多义词嵌入如何解决聚类消歧问题，提高了嵌入性能，进一步增强了 SCDV[8]。证明了基于上下文消除多义词的歧义可以更好的进行文档表示。Gupta 还表明，聚类中的稀疏性约束是有利的。进一步提高 SCDV 的文档表示能力需要进一步提高消除文本词语歧义能力。

为了弥补以上工作的缺陷，文章使用预训练的 BERT 上下文嵌入作为更稳健的语义消歧感知词语嵌入与 SCDV 软聚类相结合并调整各向异性以提高文本语义消歧的综合性能，从而更有效的进行文本表示。

2　模型架构

本文提出的模型 SCDVAB 的框架主要由四个模块组成：(1)语料库语境化；(2)调整各向异性；(3)词簇向量的形成；(4)文档表示的形成。首先，通过语料库语境化模块消除该词在语料库文档中的歧义出现，这个过程发生在语料库中的每一个独特词上；其次，通过在 BERT 模型上调整各向异性减少对文本词语语境化影响；然后，通过词簇向量形成模块将上一步获得的上下文化词嵌入聚类到 k 个划分中，进行稀疏概率分布加权获得词簇向量，这一过程发生在语料库中的每个消歧词上；最后，通过文档表示模块最终生成稀疏复合文档特征向量 $SCDV_n$。SCDVAB 文本表示模型流程图如图 1 所示。具体过程如算法 1 所示。

2.1　语料库语境化

SCDVAB 表示法第一步是语料库语境化，目的是通过单独的解释消除语料库文档中词语出现的歧义。例如，“植物是靠它的根从土壤中吸收水分”中的“水分”一词和“他说的话有很大的水分”中的“水分”一词，基于它不同的使用语境有不同的含义。给定一个词语 W 及其在语料库文本中的所有出现的语境形式 $W_1, W_2, \dots, W_n$，对每个 W_i 利用预训练语言的 BERT 得到其上下文化嵌入表示 b_{wi}。将词语消歧问题视为上下文化词语向量的局部聚类问题[10]。对通过预训练 BERT 模型获得的上下文化词语嵌入 b_{wi} 进行聚类。使用 k 均值聚类将语义消歧词向量 b_{wi} 聚类到语料库 V 中词的 k 个划分中，其中 k 表示语料库所有文本中词语 W 的全部可能的解释。在上下文语义空间中，余弦距离能够反映方向上的差异，故可使用文本词语间的余弦距离作为聚类度量。

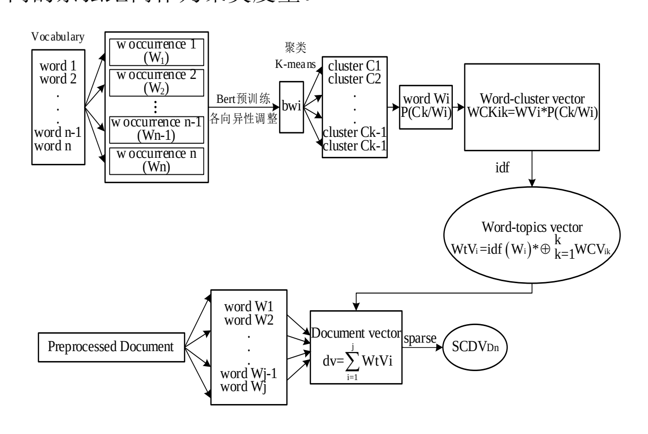


图 1　SCDVAB 模型流程图

Fig. 1　SCDVAB model flow chart

算法 1　SCDVAB(SCDV+Anisotropy+BERT)算法

输入： 文档 D_n，$n=1\dots N$。

输出： 文档向量 $SCDVD_n$，$n=1\dots N$。

对于每个 W_i，运用 BERT 模型计算语境化嵌入表示 b_{wi}；

计算 idf 值：$idf(W_i)$；

计算

$$SelfSim_\ell(W)=\frac{1}{n^2-n}\sum_j\sum_{k\neq j}\cos\left(f_\ell(s_j,i_j),f_\ell(s_k,i_k)\right)$$

$$IntraSim_\ell(s)=\frac{1}{n}\sum_i\cos(s_\ell,f_\ell(s,i))$$

其中 $s_\ell=\frac{1}{n}\sum_i f_\ell(s,i)$

$$MEV_\ell(W)=\frac{\sigma_1^2}{\sum_i\sigma_i^2}；$$

基于 K-means 模型对 b_{wi} 聚类，形成 K 类；

将 C_{w1}，C_{w2}，…，C_{wk} 分别作为 K 类的中心节点；

基于词语 W_i 和计算类 C_k，计算其条件依赖概率 $P(C_k|W_i)$；

```
for 词汇表 V 中每一个词语 W_i
for 每一个 C_k
计算 WCK_ik = WV_i * P(C_k|W_i)；
end
计算 Wtv_i = idf(W_i) * ⊕_{k=1}^{k} WCv_i k；
end
for n∈(1..N)do
初始化文本向量 dVD_n = 0；
for 词语 W_i in Dn
计算 dv = Σ_{i=1}^{p} Wtv_i；
end
计算 SCDVD_n = make-sparse(dVD_n)；
end
```

2.2　词簇向量的形成

设 $C_{w1}, C_{w2}, \cdots, C_{wk}$ 为对词语 W 进行 k 均值聚类后得到的 k 个聚类质心。将 k 个质心表示视为词语 W 的 k 个意义的多义词表示。在对语料库中出现的每个词语 W 进行聚类后，计算 BERT 表示和质心嵌入(即 $C_{w1}, C_{w2}, \cdots, C_{wk}$)之间的余弦相似度来执行上下文化的词义消歧，找到最近的聚类质心 j，即该事件的词义作为该词语 W 出现的上下文消歧词语嵌入。指定嵌入 C_{wj} 的最近邻聚类质心作为该词 W 出现的语境化消歧词嵌入。对所有出现的词语 W 重复上述过程，获得最终意义的上下文消歧词语嵌入。词语 W 的每一个语境化嵌入都充当了消除歧义的词语向量。

2.3 调整各向异性

调整各向异性的过程使用三种不同的度量标准来衡量一个词的上下文表示方式：自相似性、句内相似性和最大可解释方差[11, 12]。对于自相似性和句内相似性，基线来自不同上下文的均匀随机抽样词语表示之间的平均余弦相似性。对于最大可解释方差(MEV)，通过计算由均匀随机抽样词语表示的第一主成分解释的方差比例，并从原始 MEV 中减去该比例。使用 BERT 最后一层进行词语嵌入[13]。这里的自相似性是指 n 个唯一上下文中上下文化表示之间的平均余弦相似性。

$$SelfSim_\ell(w)=\frac{1}{n^2-n}\sum_j\sum_{k\neq j}\cos\left(f_\ell(s_j,i_j),f_\ell(s_k,i_k)\right) \quad (1)$$

其中，$f_\ell(s,i)$ 是一个将 $s[i]$ 映射到模型 f 的 ℓ 层中表示的函数。

词语 w 越语境化，自相似性越低。一个句子的句内相似度是它的词表示和句子向量之间的平均余弦相似度，也就是这些词向量的平均值。

$$IntraSim_\ell(s)=\frac{1}{n}\sum_i\cos(s_\ell,f_\ell(s,i))$$

$$\text{where}\quad s_\ell=\frac{1}{n}\sum_i f_\ell(s,i) \quad (2)$$

最大可解释方差是 w 对给定层的语境化表示的方差比例，可以用第一主成分来解释。说明静态嵌入可以在多大程度上替代词语的上下文表示。

$$MEV_\ell(w)=\frac{\sigma_1^2}{\sum_i\sigma_i^2} \quad (3)$$

其中，$[f_\ell(s_1,i_1)\ldots f_\ell(s_n,i_n)]$ 是 w 的事件矩阵，σ 是矩阵的奇异值。

为了调整各向异性的影响，使用三条各向异性基线，每一条基线对应上下文度量。对于自相似性和句内相似性，基线是来自不同上下文的均匀随机抽样词语表示之间的平均余弦相似性。给定层中的词语表示越各向异性，该基线越接近 1。对于最大可解释方差，基线是由第一主成分解释的均匀随机抽样的词语表示中的方差比例。给定层中的表示越各向异性，该基线越接近 1。从每个度量值中减去其各自值基线，获得各向异性调整的同期性度量。原始度量和基线都是使用 1K 均匀随机抽样的词语表示进行估计的。

$$Baseline(f_\ell)=E_{x,y\sim U(o)}[\cos(f_\ell(x),f_\ell(y))]$$

$$SelfSim_\ell^*(w)=SelfSim_\ell(w)-Baseline(f_\ell) \quad (4)$$

其中，o 是所有词语出现的集合。上下文化表示通常在较高的层中更具各向异性[14]。上下文各向异性在不同的模型中的表现也是不同的。BERT 层越高，平均自相似性越低。相反地，层次越高，上下文化表示就越具体[15]。同一个词在不同语境中的表示仍然比两个不同词的表示具有更大的余弦相似性，这种自相似性在上层要低得多。语境化模型的上层会产生更为特定的语境表示，很像 LSTM 的上层如何生成更多特定于任务的表示。

2.4 文档表示

针对预训练 BERT 获得的每个词语 w_i 的词向量 wv_i，计算 idf 值 $idf(w_i)$，$i=1\ldots|V|$。其中 $|V|$ 是词汇量。通过引入软聚类确保每个词都以一定的概率 $P(C_k|w_i)$ 属于每个聚类类别。

通过贝叶斯规则计算给定主题词和给定词语 w_j 的概率 $P(C_k|w_i)$。其中：

$$P(w_k|c_i)=\frac{P(c_i|w_k)P(w_k)}{P(c_i)} \quad (5)$$

$$P(c_i)=\sum_{i=1}^{k}P(c_i|w_k)P(w_k) \quad (6)$$

$$P(w_k)=\frac{\#(w_k)}{\sum_{i=1}^{V}\#(w_i)} \quad (7)$$

对于词汇表中每一个词语 w_i 及每一个聚类 C_k，

$$wck_{ik}=wv_i*P(c_k|w_i) \quad (8)$$

对于每个词语 w_i，通过加权词语在第 k 个聚类中的概率分布 $P(C_k|w_i)$，创建 k 个不同的 d 维词语聚类向量 WCV_{ik}。然后，将所有 k 个词聚类向量 WCV_{ik} 连接到一个 K×d 维嵌入中，并使用 W_i 的逆文档频率即 idf 对其进行加权，形成一个上下文化的词主题向量 WtV_i。

$$wtv_i=idf(w_i)\times\oplus_{k=1}^{K}wcv_{ik} \quad (9)$$

其中，$\oplus$ 是串联的意思。

初始化文档向量 $dVD_n=0$，$n\in(1..N)$。最后，对于文档 D_n 中出现的所有词语，将它们的词主题向量 WtV_i 相加获得文档向量 dVD_n。

$$dV=\sum_{i=1}^{j}wtv_i \quad (10)$$

对向量进行归一化，dVD_n 中的大多数值都非常接近于零[16]。通过将绝对值接近阈值的属性值归零，使文档向量 dVD_n 稀疏，从而生成稀疏复合文档向量 $SCDVD_n$。

$$SCDVD_n=make-sparse(dVD_n) \quad (11)$$

3 实验与分析

为了评估 SCDVAB 算法的综合性能，首先对算法的嵌入精确性和其他最新上下文嵌入技术进行对比，并且在概念匹配和语义文本相似度任务上进行了实验。

3.1 实验环境

算法的实验环境如表 1 所示。

表 1　实验环境

Tab. 1　Experimental environment

实验环境	环境配置	实验环境	环境配置
CPU	Intel® Core™i7-10710U	编程语言	Python3.7.11
操作系统	Windows10	开发工具	Pycharm
内存	32GB	深度学习框架	Tensorflow 2.4.1

3.2 数据集和基线

为了分析语境化的词语表示，需要输入句子到预先训练好的模型中。在 4 个广泛使用并且公开的分类数据集上进行了实验比较精确性：(1)Amazon 数据集，有 4 个类别，8000 条文本；(2)Classic 数据集，有 4 个类别，7095 条文本；(3)20NG 数据集，是新闻组文本数据集，有 20 个类别，每个类别样本数目相同，一共包含 18846 篇文本；(4)Twitter 数据集，有 3 个类别，3115 条文本。实验将 doc2vecc，idf 加权的 word2vec，BERT，SCDV+word2vec，SCDV+BERT(加权平均值)，SCDV+BERT 设为对比基线。特别地，设置 SCDV+BERT(加权平均值)为基线，是为了分析基于词义消歧的词向量能够更有效地捕捉词的多重含义。设置 SCDV+BERT 基线，目的是分析减少了各向异性的影响。使用 k=6 配合各向异性调整。基线取自 Gupta et al,2020 论文的实验部分[17]。

概念匹配任务是将概念与相关项目联系起来。概念匹配数据集包括来自下一代科学标准 3 (NGSS)的 53 个独特概念的 537 对项目和概念，以及来自 Science Buddies 的 230 个独特项目。实验与 TF-IDF 加权向量、SCDV+Word2Vec 预训练的 BERT 基线之间的余弦相似度进行对比。基线取自 2020 年 Zhang 和 Danescu-Niculescu-Mizil 的实验部分[18]。

句子相似性任务是计算两个文本在语义层面的相似性，实验的输入数据来自涉及 2012-2016 年间的 27 项语义文本相似性(STS)任务[19]。数据集中每年有 4 到 6 项 STS 任务，详细任务见表 2。使用这些数据集是因为它们包含相同词语出现在不同上下文中的句子。在所有的数据集中，每一个词语都有多个多义词。基线取自 Perone et al,2018[20]、Devlin et al,2019[21]以及 Gupta et al,2020[17]的实验部分。

表 2 STS 任务

Tab. 2 STS task

STS12	STS13	STS14	STS15	STS16
MSRpar	headline	deftforum	answer-forums	headlines
MSRvid	OnWN	deft news	answers-students	plagiarism
SMT-eur	FNWN	headline	belief	postediting
OnWN	SMT	images	headline	ans-ans
SMT-news		OnWN	images	ques-ques
		Tweet news		

3.3 实验设置

使用 BERT 无基础预训练模型获得词语嵌入，并使用 K-means 对给定词语进行上下文聚类。为了简单起见，实验对所有的数据使用了 0.8 的相似性阈值(τ)，这导致每个词都有多个多义词表示。统计相似程度的分布，其中，实验不考虑出现在不到 5 个独特上下文中的词语。训练集和测试集按八二比例划分，对于 SCDV，将词语嵌入的维度设置为 200，设置 k=6 进行各向异性调整，使用 5 倍交叉验证来调整 SCDV 的稀疏阈值。

3.4 实验结果分析

表 3 为 SCDVAB 与其他基线模型在 4 个数据集上的精确性表现，实验结果为各模型训练 10 次的平均值。从表 2 实验结果可知，SCDVAB 模型在所有数据集上比其他的上下文语境文本表示方法效果都更为出色。

表 3 SCDVAB 与各基线精确性对比

Tab. 3 Comparison of SCDVAB and baseline accuracy

Embedding	20NG	Amazon	Classic	Twitter
Doc2vecC	78.20	91.10	96.60	71.00
Word2vec(idf 加权)	81.70	93.90	95.20	72.00
BERT	64.78	91.04	95.63	66.64
SCDV+word2vec	84.87	93.84	96.90	74.17
BERT(加权平均)+SCDV	84.88	94.59	95.62	72.98
BERT+SCDV	86.07	94.15	97.81	75.97
SCDVAB	86.92	95.87	99.01	77.03

通过表 3 实验结果分析，语境化的 BERT+SCDV 比加权平均的 BERT+SCDV 表现更好。词向量的简单加权通常能够产生有效的句子表示，但表示包含多个句子的长文本时，相比基于词义消歧的词向量效果要差。这是因为较长句子的文本可能包含大量不同话题的词语。实验结果表明基于词义消歧的词向量能够捕捉到词的多重含义，证明了语义消除歧义的贡献。其次，SCDVAB 模型相比 BERT+SCDV 模型的精确度分别高了 0.85%、1.72%、1.2%和 1.06%，证明了调整各向异性的优势影响。SCDVAB 模型的性能优于 BERT(加权平均)+SCDV,这表明 SCDVAB 基于词义消歧的词向量能够有效地捕捉多义词，调整各向异性能够提升语境化表示能力，更符合语料库语境。

表 4 概念匹配精确率及 F1 值对比

Tab. 4 Comparison of concept matching accuracy and F1 value

Embedding	Accuracy	F1
TF-IDF	53.8	70.0
BERT	54.7	70.6
Word2vec+SCDV	53.6	70.0
BERT+SCDV	57.1	73.8
SCDVAB	58.9	74.6

基于表 4 观察各模型的性能表现，SCDVAB 模型在精确率和 F1 值上分别比预训练的 BERT 模型和 Word2Vec+SCDV 模型分别高出 4.2%、4%和 5.3%、4.6%。对比 BERT+SCDV 模型在精确率和 F1 值上分别高出了 1.8%和 0.8%，证明了 SCDVAB 模型在概念匹配任务上的优越性，侧面体现了 SCDVAB 模型在解决文本词语的歧义性以及准确界定词语在特定上下文语境中的语义特征的性能上的优势。

表 5 SCDVAB 与各种 STS 任务的最新嵌入技术对比

Tab. 5 Comparison of the latest embedding technologies between SCDVAB and various STS tasks

Embedding	Y12	Y13	Y14	Y15	Y16	Avg.
ELMO orig+all	55	51	63	69	64	60.4
ELMO orig+top	54	49	62	67	63	59
BERT	53	67	62	73	67	64.4
P-mean	54	52	63	66	67	60.4
fastText	58	58	65	68	64	62.6
Skip Thoughts	41	29	40	46	52	41.6
PSIF+PSL	65.7	64.0	74.8	77.3	73.7	71.1
u-SIF+PSL	65.8	65.2	75.9	77.6	72.3	71.4
Word2vec+SCDV	64.1	63.9	73.0	76.9	77.3	71.0
BERT+SCDV	64.7	64.0	75.4	77.1	77.3	70.9
SCDVAB	66.8	64.1	77.3	78.0	74.6	72.2

表 5 展示了 SCDVAB 模型与各种最新嵌入技术的比较。实验的数据为皮尔逊相关系数乘以 100。观察各模型在数据集上的性能表现，SCDVAB 模型显著优于其他基线模型，证明了改进模型的有效性。根据实验结果观察到，通过结合 SCDV 的算法模型比其他算法产生更好的性能。这种情况的主要原因是，SCDV 通过对预先训练的词向量进行软稀疏聚类，进一步将表示性能从句子扩展到文本，证明了 SCDVAB 利用 SCDV 的优越性。与 Word2vec+SCDV 相比，由于考虑了词义消歧以及语境化表示能力，SCDVAB 显著提高了 Pearson 的分数。BERT+SCDV 相比 Word2vec+SCDV 略有改进，但相比改进的 SCDVAB 还是略有逊色的。这是因为 SCDAB 模型考虑了调整各向异性对于 BERT 词义消歧的影响。

为了验证 SCDVAB 模型对比其他模型的性能优越性，下面展示了 STS12 数据集中 MSRvid 任务中的的几条样本相似性用于部分实验结果的可视化和分析，样本描述如表 6 所示，表中数据已做标准化处理。

表 6 STS12 MSRvid 数据集相似性样本实例对

Tab. 6 STS12 msrvid dataset similarity example pair

sentence 1	sentence 2	GT	PSIF	BERT+SCDV	SCDVAB
Runners race around a track.	Runners compete in a race	0.64	0.7453	0.6933	0.6418
A man is riding a motorcycle.	A woman is riding a horse.	0.15	0.2725	0.164	0.160
People are playing baseball.	The cricket player hit the ball.	0.1	0.2371	0.12	0.0973
A animated airplane is landing.	A plane is landing.	0.56	0.6338	0.7206	0.5773

观察表 6 实验结果发现，在 SCDVAB 模型上得到的相似性分数对比其他模型都更接近给定的相似性，证明了改进模型在计算两个文本在语义层面的相似性上的优越性。

在表 7 探讨了几个模型在 STS16 上文本相似性任务的实验结果，用于进一步验证 SCDVAB 的改进对于性能的提升。

根据表 7 的实验结果可以看出，在 STS16 任务中改进模型在所有数据集上都优于其他算法，证明了 SCDVAB 模型的优越性。对比各模型在数据集上的表现，PSIF+PSL 模型效果优于 skip thought 模型。这是因为 P-SIF 从文本中学习特定于

主题的向量，考虑了文本主题结构利用了分区平均技术。而 skip thoughts 模型结构借助 skip-gram 思想，缺少考虑词语在特定上下文语境中的语义特征。而 BERT+SCDV 模型与 P-SIF+PSL 模型相比效果相差不大，但略微差一些。猜测原因可能为，未做改进的 BERT 对于文本长度有限制，而 P-SIF+PSL 模型是针对长文本的分区平均算法更有针对性。SCDVAB 模型对比 BERT+SCDV 模型性能有所提升，进一步体现了考虑各向异性的重要性。

表 7 各模型 STS16 上文本相似性任务实验结果

Tab. 7 Experimental results on textual similarity tasks on STS 16

Tasks	Skip thoughts	PSIF+PSL	BERT+SCDV	SCDVAB
headlines	51.12	75.6	74.7	76.2
plagiarism	66.77	81.6	81.3	82.3
Post editing	69.95	83.7	83.6	84.7
ans-ans	28.83	60.2	60.1	61.6
ques-ques	40.66	67.2	66.9	68.2
STS16	52	73.7	73.3	74.6

4 结束语

考虑文本表示时需要解决的词语歧义性问题，以及词语在特定上下文语境中的语义特征问题，本文提出了文本语义消歧的 SCDVAB 算法模型。通过预先训练的 BERT 上下文化，并减少各向异性的影响来增强稀疏文本表示(SCDV)，为上下文文档表示提供了一个更高效、更准确的文本表示方法。

基于各向异性调整之后的 BERT 语义消歧词向量，运用 SCDV 转换为文本的特征向量，可准确表示词语在特定上下文语境中的语义特征，具有较强的实际意义。实验结果表明，SCDVAB 模型优于其他无监督方法，在文本语义消歧的综合性能上更出色。相关模型可有效提高多主题长文本表示、多场景文本概念消歧以及抽取式阅读理解等文本表示相关任务的效率。

参考文献：

[1] Mekala D, Zhang X, Shang J. META: Metadata-empowered weak supervision for text classification [C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. [S. I.] : Association for Computational Linguistics, 2020: 8351-8361.

[2] 刘欢，张智雄，王宇飞. BERT 模型的主要优化改进方法研究综述 [J]. 数据分析与知识发现，2021, 5 (1): 3-15. (Liu Huan, Zhang Zhixiong, Wang Yufei. Review on the main optimization and improvement methods of Bert model [J]. Data Analysis and Knowledge Discovery, 2021, 5 (1): 3-15.)

[3] Gong H, Sakakini T, Bhat S, *et al.* Document similarity for texts of varying lengths via hidden topics [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2018: 2341-2351.

[4] 刘胜杰，许亮. 基于词嵌入技术的文本表示研究现状综述 [J]. 现代计算机，2020, 673 (1): 40-43. (Liu Shengjie, Xu Liang. Summary of research status of text representation based on word embedding technology [J]. Modern Computer, 2020, 673 (1): 40-43.)

[5] 焦芬芬. 基于概念和语义相似度的文本聚类算法 [J]. 计算机工程与应用，2012, 48 (18): 136-141. (Jiao Fenfen. Text clustering algorithm based on concept and semantic similarity [J]. Computer Engineering and Application, 2012, 48 (18): 136-141.)

[6] 王瑞琴，孔繁胜. 无监督词义消歧研究 [J]. 软件学报，2009, 20 (8): 2138-2152. (Wang Ruiqin, Kong Fansheng. Research on unsupervised word sense disambiguation [J]. Journal of Software, 2009, 20 (8): 2138-2152.)

[7] Mekala D, Gupta V, Paranjape B, *et al.* SCDV: Sparse Composite Document Vectors using soft clustering over distributional representations [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 659-669.

[8] Gupta V, Saw A, Nokhiz P, *et al.* Improving document classification with multi-sense embeddings. [C]/ Proceedings of the European Conference on Artificial Intelligence, 2020. (2020-11) [2022-2-20]. http://10. 48550/arXiv. 1911. 07918.

[9] Ethayarajh K. How contextual are contextualized word representations?comparing the geometry of BERT, ELMo, and GPT-2 embeddings. [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics2019: 55–65.

[10] Matthew P, Mark N, Mohit I, *et al.* Deep contextualized word representations [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. [S. I.] : Human Language Technologies, 2018 (1): 2227–2237.

[11] Yosinski J, Clune J, Bengio Y, *et al.* How transferable are features in deep neural networks? [J]. Advances in Neural Information Processing Systems. 2014 (11): 3320–3328.

[12] Bhatia K, Jain H, Kar P, *et al.* Sparse local embeddings for extreme multi-label classification [J]. Advances in Neural Information Processing Systems, 2015 (1): 730–738.

[13] YuMeng, Shen Jiaming, Zhang Chao, *et al.* Weakly-supervised hierarchical text classification [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019 (33): 6826–6833.

[14] 叶雪梅，毛雪岷，夏锦春，等. 文本分类 TF-IDF 算法的改进研究 [J]. 计算机工程与应用，2019, 55 (2): 104-111. (Ye Xuemei, Mao Xuemin, Xia Jinchun, *et al.* Improvement of TF-IDF algorithm for text classification [J]. Computer Engineering and Application, 2019, 55 (2): 104-111.)

[15] 戴洪涛，侯开虎，周洲，等. 基于 VCK-vector 模型的词义消歧方法 [J]. 软件，2020, 41 (2): 134-140. (Dai Hongtao, Hou Kaihu, Zhou Zhou, *et al.* Word sense disambiguation method based on VCK vector model [J]. Software, 2020, 41 (2): 134-140.)

[16] 王瑞，李弼程，杜文倩. 基于上下文词向量和主题模型的实体消歧方法 [J]. 中文信息学报，2019, 33 (11): 46-56. (Wang Rui, Li Bicheng, Du Wenqian. Entity disambiguation method based on context word vector and topic model [J]. Chinese Journal of Information Technology, 2019, 33 (11): 46-56.)

[17] Gupta V. , Saw A. , Nokhiz P. , *et al.* P-SIF: Document embedding using partition averaging [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34 (5): 7863-7870.

[18] Zhang J, Danescu-Niculescu C. Balancing objectives in counseling conversations: Advancing forwards or looking backwards [J]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguists, 2020: 5276-5289.

[19] Kim, H K, Kim H, Cho S. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation [J]. Neurocomputing, 2017 (266): 336–352.

[20] Perone, C. S. , Silveira, R. , Paula, T. S. Evaluation of sentence embeddings in downstream and linguistic probing tasks [J]. 2018. arXiv preprint arXiv: 1806. 06259.

[21] Devlin, J. , Chang, M-W. , Lee, K. , *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding [C/OL]. NAACL, 2018. (2018-10-11) [2022-2-20]. https://arxiv. org/abs/1810. 04805.